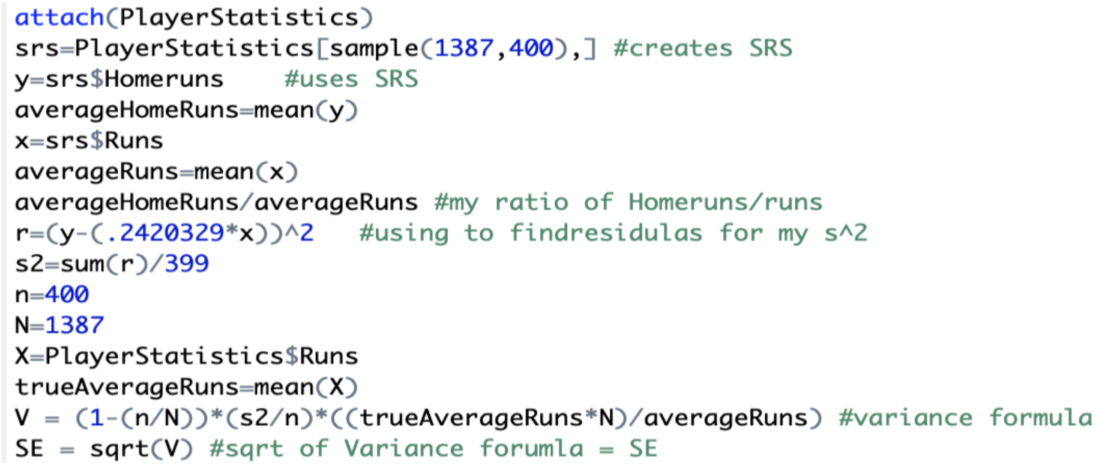
Spencer Palladino
Sampling Methods
Test 2

I hereby state that I have not had any form of interaction with any other person in answering this test.

1.

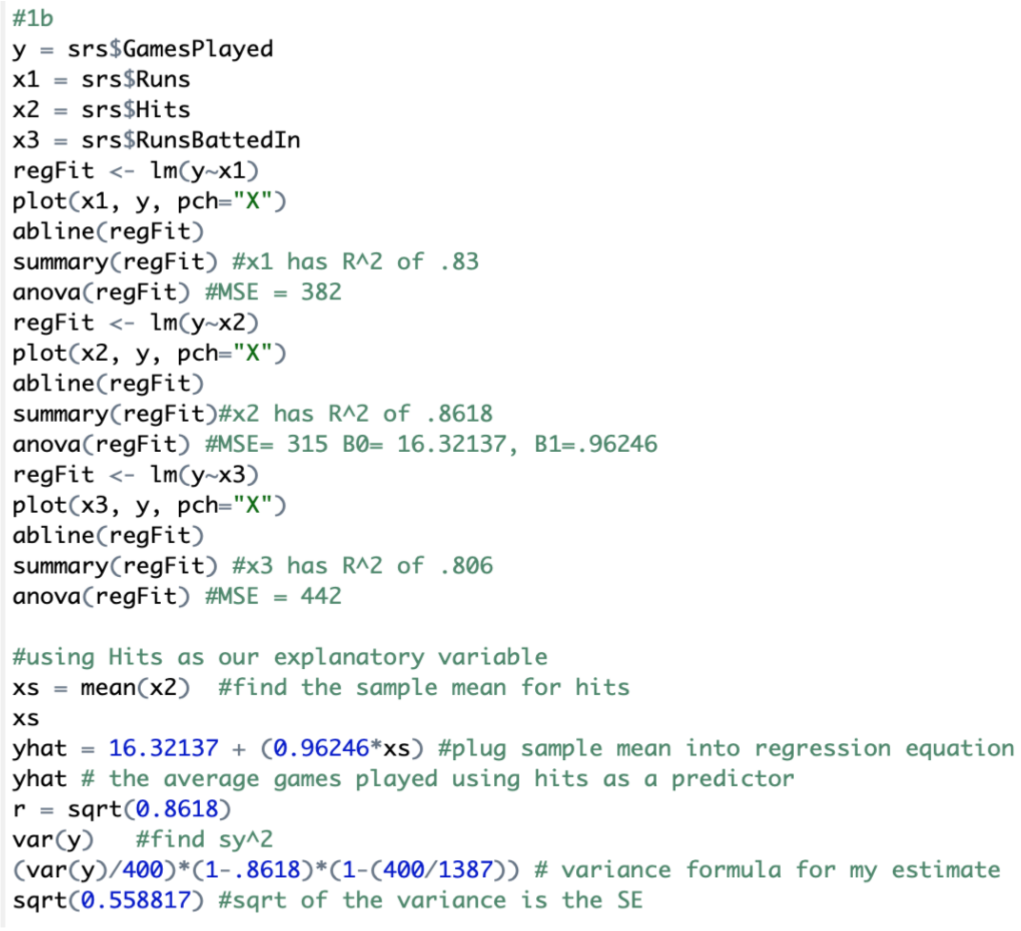
a) Homeruns/Runs ratio for my sample of 400 = 0.242 and my standard error is 4.567.



```
attach(PlayerStatistics)
srs=PlayerStatistics[sample(1387,400),] #creates SRS
y=srs$Homeruns     #uses SRS
averageHomeRuns=mean(y)
x=srs$Runs
averageRuns=mean(x)
averageHomeRuns/averageRuns #my ratio of Homeruns/runs
r=(y-(.2420329*x))^2   #using to findresidulas for my s^2
s2=sum(r)/399
n=400
N=1387
X=PlayerStatistics$Runs
trueAverageRuns=mean(X)
V = (1-(n/N))*(s2/n)*((trueAverageRuns*N)/averageRuns) #variance formula
SE = sqrt(V) #sqrt of Variance forumla = SE
```

b) regression estimate using our sample for average Games Played = 43.9175 games using Hits as my explanatory variable. My SE for the regression estimate using my sample is 0.7475406.

I used hits as my explanatory variable because after running an ANOVA test and regression summary with each variable, I found that the relationship between the variable Hits and Games played resulted in the highest R^2 value of .8618 and the lowest variance (315). Thus, we are able to explain the most amount of variability in the model using Hits as a predictor.



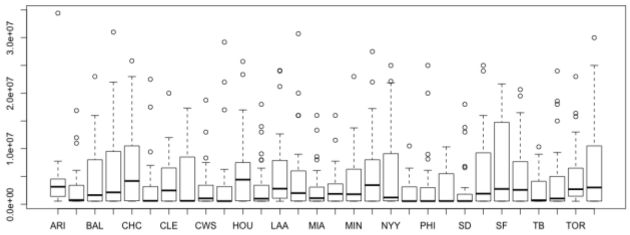
```
#1b
y = srs$GamesPlayed
x1 = srs$Runs
x2 = srs$Hits
x3 = srs$RunsBattedIn
regFit <- lm(y~x1)
plot(x1, y, pch="X")
abline(regFit)
summary(regFit) #x1 has R^2 of .83
anova(regFit) #MSE = 382
regFit <- lm(y~x2)
plot(x2, y, pch="X")
abline(regFit)
summary(regFit)#x2 has R^2 of .8618
anova(regFit) #MSE= 315 B0= 16.32137, B1=.96246
regFit <- lm(y~x3)
plot(x3, y, pch="X")
abline(regFit)
summary(regFit) #x3 has R^2 of .806
anova(regFit) #MSE = 442

#using Hits as our explanatory variable
xs = mean(x2)  #find the sample mean for hits
xs
yhat = 16.32137 + (0.96246*xs) #plug sample mean into regression equation
yhat # the average games played using hits as a predictor
r = sqrt(0.8618)
var(y)   #find sy^2
(var(y)/400)*(1-.8618)*(1-(400/1387)) # variance formula for my estimate
sqrt(0.558817) #sqrt of the variance is the SE
```

2.

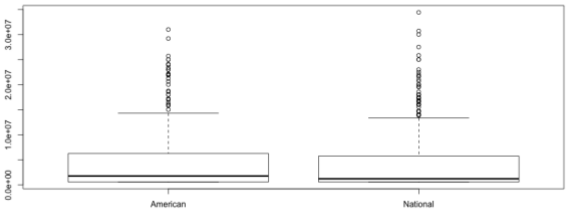
a) Clustering is a better team variable than League. Team has an R^2 adjusted of 0.02338 while league has an R^2 adjusted of -0.000633. A positive R^2 adjusted value means that there is a clustering effect in the data. R^2 adjusted being negative is a good initial sign that clustering would be more effecting than an SRS as elements in the same cluster may be more variable than randomly selected elements in the whole population.

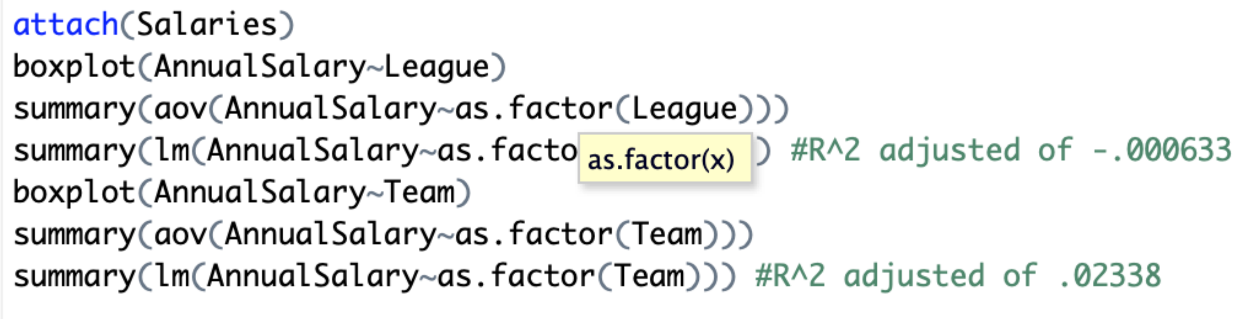
Looking at the box plots of the two different clustering variables, the mean and standard deviation of the two clusters when using league as a variable are nearly identical. When using Team as a clustering variable, there is much more obvious changes in the mean and standard deviations for each team (meaning there is a higher variability between psus which is bad for clustering). Thus, clustering would be more effective when league is the clustering variable as it will be far more useful than when using team as our variable.

Team Box Pot



League Box Plot

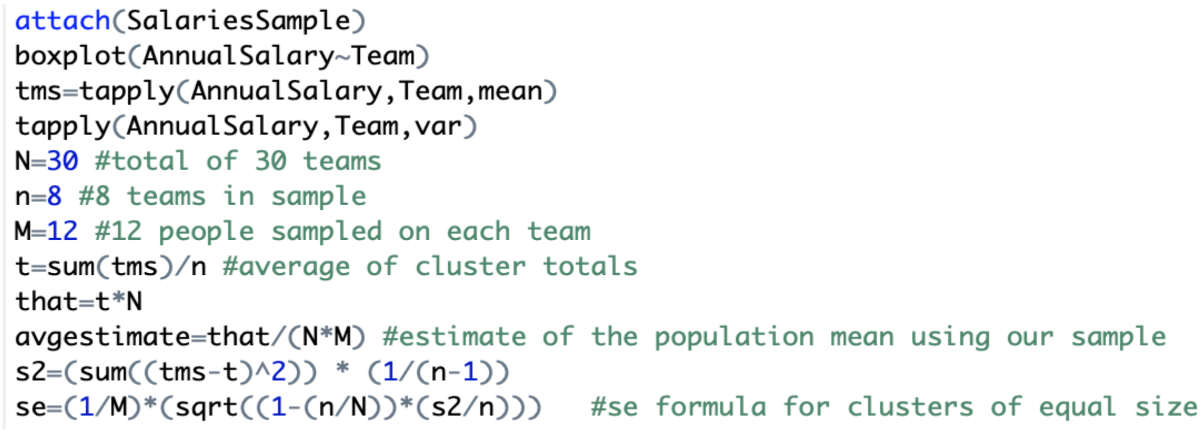




```
attach(Salaries)
boxplot(AnnualSalary~League)
summary(aov(AnnualSalary~as.factor(League)))
summary(lm(AnnualSalary~as.facto as.factor(x) ) #R^2 adjusted of -.000633
boxplot(AnnualSalary~Team)
summary(aov(AnnualSalary~as.factor(Team)))
summary(lm(AnnualSalary~as.factor(Team))) #R^2 adjusted of .02338
```

b)

estimate of true average annual salary for MLB player in 2018 = $426,911.9 with a standard error=$52,837.0



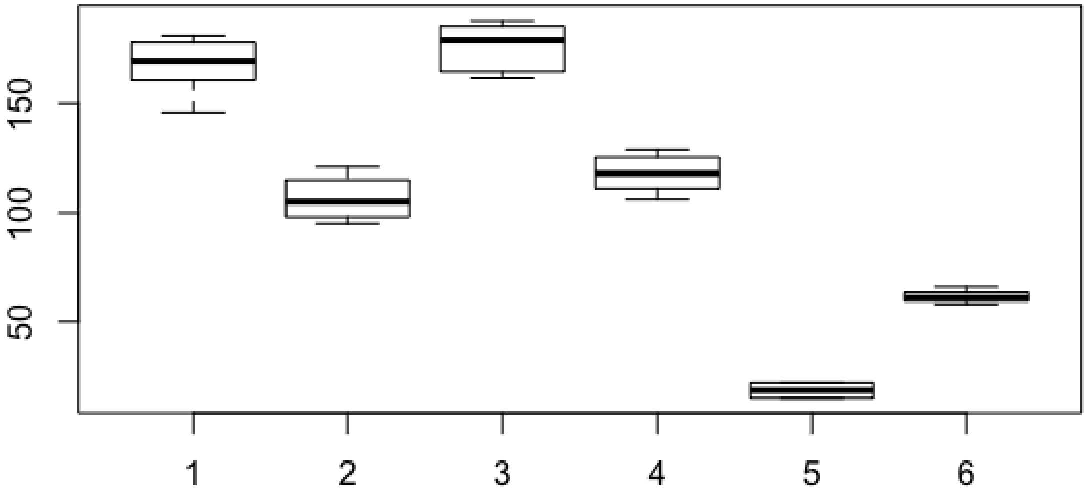
```
attach(SalariesSample)
boxplot(AnnualSalary~Team)
tms=tapply(AnnualSalary,Team,mean)
tapply(AnnualSalary,Team,var)
N=30 #total of 30 teams
n=8 #8 teams in sample
M=12 #12 people sampled on each team
t=sum(tms)/n #average of cluster totals
that=t*N
avgestimate=that/(N*M) #estimate of the population mean using our sample
s2=(sum((tms-t)^2)) * (1/(n-1))
se=(1/M)*(sqrt((1-(n/N))*(s2/n)))   #se formula for clusters of equal size
```

c) If we sampled less teams but more players from each team and continued to assume that each team has the same number of players per sample, n would decrease to be less than 8 (refer to code in part b) and M would increase to be total number of players on each team. A maximum variability would occur when the minimum number of teams are selected as clusters which would be 2 and those teams have the largest difference in average annual salary (Chicago Cubs have the highest average annual salary and the Oakland A's have the lowest). A zero-variability situation would occur when there is only one team in the sample as there are no other psus. The lowest variability will occur when teams being sampled are very close to each other in mean annual salary.

3.

a) The situation described is a two-stage cluster sample in which the market research firm selects 6 random cities (psus) from the population and then randomly selects a number of stores in those cities (ssus) to send the candies. Given that a cluster sample is defined as a sampling design with observations grouped in clusters (psus), I can confidently say that this sample is using clustering as the 6 random cities are the psus.

b) A sample clustering design is not ideal in this situation as the psus are pretty homogeneous. There is a ton of variability in between the clusters, and the data has an R^2 adjusted value of 0.9539. This is in agreement with my claim that most of the variability is in between clusters. This means that little new information will be shown by sampling with clusters which means and SRS may be a better idea for this experiment.


50
100
150
1
2
3
4
5
6